DOCUMENT RESUME

ED 447 196 TM 032 085

AUTHOR	Cizek, Gregory J.
TITLE	Factors Affecting Linkage of the Voluntary National Test and the Proposed Voluntary National Tests.
PUB DATE	2000-04-00
NOTE	21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Achievement; Elementary Secondary Education; Equated Scores; *National Competency Tests; Standardized Tests; *Test Use
IDENTIFIERS	Linking Educational Assessments; Linking Metrics; *National Assessment of Educational Progress; *Voluntary National Tests

ABSTRACT

Issues involved in linking the National Assessment of Educational Progress (NAEP) to the proposed Voluntary National Tests (VNTs) are discussed. Linkage is used to refer to procedures intended to permit scores from two different tests that are designed to measure the same variable to be expressed on the same scale. There are substantial differences between NAEP and the VNT that present serious challenges to linking the VNT to NAEP. The single greatest consideration in evaluating the potential for score scales to be linked is that of construct equivalence. Allied with the notion of construct equivalence are questions of the purposes of the assessments, administration conditions, and the stakes, visibility, and motivation. The greatest impact on overall construct equivalence is the extent to which content covered on the proposed VNTs can be viewed as consistent with that covered by the respective NAEP tests. Achievement levels, reporting methods, interpretations, and audiences must also be considered. The technical problems are serious enough, and the weight of policy considerations and uncertainty about how a VNT will affect NAEP are also worth contemplating. There are policy issues that should be addressed before considering linking methods. (Contains 1 table and 11 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Factors Affecting Linkage of the Voluntary National Test and the Proposed Voluntary National Tests

ED 447 196

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

G. J. Cizek

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM032085

Gregory J. Cizek
Associate Professor of Educational Measurement and Evaluation
110 Peabody Hall
School of Education, CB 3500
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3500
Tel: 919-843-7876
Email: cizek @unc.edu

Paper presented at the annual meeting of the National Council on Measurement in Education, April 2000, New Orleans, LA

BEST COPY AVAILABLE

Factors Affecting Linkage of the Voluntary National Test and the Proposed Voluntary National Tests

Ok, I admit it. As you can probably discern from the title of this paper, there will be something of a bait and switch going on here. My presentation today will be true to the title of the session, introducing the factors that affect linkages generally, and factors that make linking the National Assessment of Educational Progress (NAEP) to darn near anything particularly difficult. As most of us probably know, the specific matter under consideration is how to link the NAEP to the proposed Voluntary National Tests (VNTs) in reading and mathematics.

I am supposed to provide an overview of the technical/psychometric issues involved, and I will. However, let me begin by expressing what is only an honest self-evaluation: As I think about the technical linking expertise represented on this panel, I have to ask myself the obvious question: "What the heck am I doing here? There is nothing I will say in the first portion of my presentation that my colleagues could not have summarized more concisely, or that they will not in a few moments elaborate upon more accurately and with more authority.

On the other hand, allow me to foreshadow my later remarks by noting something else: There is nothing in what I will say in the second portion of my presentation that my colleagues would, perhaps, *want* to say. I respect them for that, too, and I accept that I may be mentioned in the same breath with snippets of adages like "where angels fear to tread." I admit, too, to being less informed than many--even all--members of this panel vis-à-vis the political issues swirling around this NAEP/VNT development process. Another proverb reminds us that "ignorance is bliss," and just look at me: I don't work for the Congress, AIR, NCES,

etc., and I am the happiest guy here today. (In fact, depending on how the presentation turns out, I may not work for NAGB anymore either!)

So what *am* I doing here? In the next few minutes, I hope to present the factors affecting linking NAEP and VNT in the broadest possible way, and in doing so make a contribution as a person who combines interests in technical issues with an eye toward educational policy issues. With that bit of introduction, let us turn to a brief overview of factors that affect the linkage of NAEP and VNT.

An Overview of Linking Considerations

There are many good sources for those interested in obtaining information about the various procedures that can be used to link measures, and the contexts in which those procedures are most appropriate and yield the most satisfactory linkages. A modest overview is provided in the booklet distributed at this session (Cizek, Kenney, Kolen, Peters, & van der Linden, 1999); several other more or less extensive treatments are also available (e.g., Mislevy, 1992; National Research Council, 1999; Petersen, Kolen, & Hoover, 1989), and high-quality, data-based work on the viability of various strategies for linking educational assessments is available (see, for example, Waltman, 1997; Williams, Rosa, McLeod, Thissen, & Sanford, 1998). Much of the following is drawn from the report by Cizek, et al (1999), the full version of which is available here today.

In this paper, the term *linkage* is used to refer to procedures intended to permit scores from two different tests which are designed to measure the same variable to be expressed on the same scale. Linking methods can be used to adjust the first set of scores so as to express them on the metric of the second test; or to adjust the second set of scores so as to express them on the metric of the first; or to express both sets of scores on third, common scale.

Several desirable results can be obtained when score scales are linked. One such result is that, depending on the approach used, linking can facilitate comparisons between the performance of students

who took one test and the performance of different students who took a different test. Linking the scales of two tests can also enable certain predictions about how students would perform on one test based on knowledge of how they performed on the other test. It has been stated that "the quality of linkage hinges on how well one can infer from the performance on test B the proficiencies that test A is designed to measure" (National Research Council, 1999, p. 12). To put this idea into the current context: it has been suggested it may be of interest to express student performance on the VNT in terms of the NAEP scale--at least one advantage of this being the ready interpretability of VNT performance in terms of NAEP achievement levels. The quality of NAEP/VNT linkages will determine the extent to which accurate inferences about NAEP performance can be made on the basis of VNT performance.

Many variables affect the quality of linkages, however, especially when the measures to be linked are as sophisticated as NAEP is currently, and as sophisticated as any VNT is likely to be. As the two measures depart from strictly parallel construction, administration, and scoring conditions, these differences will degrade the quality of the linkage. *A priori*, we know that there are noticeable differences between NAEP and VNT which create serious challenges for linking. A scorecard of potential or known differences between NAEP and VNT is provided in Table 1. Many of these differences are summarized in the following paragraphs.

Insert Table 1 about here.

Constructs Measured

Perhaps the single greatest consideration in evaluating the potential for score scales to be linked is that of construct equivalence. If two measures actually assess the same construct, the full armamentarium of

the psychometrician (including, for example, equating and calibration methods) can be employed in the service of establishing meaningful relationships between the two score scales, or a single score scale with dependable interpretations of performance on either measure. To the extent that construct inequivalence exists, other forms of linking (e.g., statistical moderation, social moderation) are appropriate.

As currently configured, the NAEP measures various constructs at various grade levels. The VNT will measure reading achievement at grade 4 and mathematics achievement at grade 8. Although the frameworks for NAEP and VNT reading at grade 4 and mathematics at grade 8 are the same, the content specifications are different (e.g., there are more constructed response items and longer reading passages on NAEP). NAEP comprises a broad sampling of content, made possible by the use of matrix sampling; the VNT will use a single form that will almost surely be constrained in terms of number of items, passage length, content coverage, and so on. Construct equivalence is not limited to these issues, however, but encompasses all aspects of the two examinations that would serve to make the tests present different psychological tasks to examinees and yield differing interpretations. Thus, to some degree, all of the factors described subsequently in this paper as affecting linkages between NAEP and VNT are subsumed under the notion of construct equivalence.

Purposes of the Assessments

NAEP was designed to assess the achievement of groups (e.g., the nation, demographic groups, states in state NAEP) by using carefully drawn representative samples of examinees. There is no such thing as an individual student's score on NAEP and no individual-level decisions are made with NAEP. On the other hand, the VNT is specifically intended to assess the achievement of individuals. Individual scores will be provided, and individual classifications and decisions will likely be made on the basis of VNT scores. It is conceivable that the VNT might also be used to assess achievement of groups at any possible level of aggregation, such as schools, districts, or states. Like NAEP, the VNT could be used to inform policy

makers and educators about the achievement of different groups, and could conceivably be used in states' educational accountability systems. However, because the VNT is voluntary, the meaning of the aggregate data will likely depend on which students are actually administered the assessment. In summary, while the NAEP is used mainly to inform educators and policy makers about the achievement of groups, the VNT will be used to assess individuals and may be used to make high-stakes decisions about individuals and inform high stakes policy decisions.

Administration Conditions

National NAEP is administered by a central contractor; state-level NAEP is administered by the states. It is possible that the VNT could be administered by a single contractor, by states, by local districts, or other entities. Using a single contractor would likely lead to one desirable benefit--increased standardization of administration conditions. On the other hand, administration by states would involve more state personnel and might lead to a different desirable benefit--a greater sense of ownership by the states. As administration conditions of the VNT and NAEP diverge, however, the potential for the two instruments to measure different constructs increases and linking becomes more problematic. In fact, it is known that the agent conducting the administration affects resulting score distributions; for example, there is evidence from state-NAEP that scores are higher when it is administered under NAEP-S conditions than under the conditions used for the national NAEP. Additionally, if other, practical test administration differences such as order, context, fatigue, and practice effects (and others) have differential impact on VNT and NAEP performance, then different score distributions on the two instruments will occur even if they are perfectly linked.

Stakes, Visibility, and Motivation

If there is one thing certain about a mandated, individual student assessment that is implemented as

a policy mechanism to satisfy reform impulses--which the VNT assuredly is--we know that there will be different stakes associated with the VNT compared to the NAEP. This context will have effects of unknown magnitude--but almost certain direction--on factors such as student motivation, student anxiety, educator resistance, parental support, instructional alignment, journalistic reporting, and public interest. The VNT will be more visible than NAEP (more on this later) and--this is not surprising or improper--it will be used in ways that are more closely associated with its purpose than with the purpose of NAEP. These differences, of course, raise technical concerns. For example, if high stakes consequences are associated with the operational VNT administration that were not associated with the administration conditions that are present in a study to determine a linking function, then undesirable and impossible-to-reconcile differences in the score distributions will be observed.

Content Coverage/Equivalence

Perhaps the greatest impact on overall construct equivalence is the extent to which content covered by the proposed VNTs can be viewed as consistent with that covered by the respective NAEP tests. To investigate this congruence, a number of evidentiary sources are possible. For example, the two assessments' content frameworks and specifications would certainly be compared. Procedural evidence could also be obtained to investigate the extent to which item development methods and test form generation activities were likely to produce construct-equivalent measures. These issues are treated in substantially greater depth and with greater acumen elsewhere [see the thorough treatment of both reading and mathematics content and process issues by Kenney and Peters (Chapter 3) in Cizek et al. 1999].

It is fair to say that development procedures employed for NAEP are perhaps the most rigorous and professionally sound to be found on any large-scale assessment administered today. From the available evidence to date, it seems reasonable to conclude that developmental approaches followed and quality control procedures in place for the VNT are as parallel as can be those of NAEP given the time and political

constraints presented (see, e.g., Wise, Noeth, & Koenig, 1999). Nonetheless, *some* degree of inequivalence is almost certain. To a great degree, any success in linking scores on the VNT and NAEP will be determined by the extent to which parallelism in development and quality control are carried through to operational forms of the VNT and, perhaps more importantly, are configured so as to maintain as much commonality as possible throughout the operational lives of the programs.

Achievement Levels

One of the characteristics of NAEP considered desirable is the possibility of reporting student performance according to the familiar achievement levels, Basic, Proficient, and Advanced. As a whole, the NAEP development process supports inferences about the kinds of knowledge and skills that students categorized according to these levels possess. Substantial procedural and content validity information has been amassed to permit such inferences, and a solid interpretive foundation exists to aid users of NAEP in making warranted, accurate inferences to the greatest extent possible.

One of the goals of the VNT is to take advantage of the existing NAEP achievement levels and interpretive structures, by reporting VNT performance in terms of NAEP achievement levels, thereby permitting parallel inferences about knowledge and skills for students classified in similar ways. Obviously, the stronger the linkage between NAEP and the VNT, the more confident the assertion that students similarly categorized do indeed posses similar constellations of knowledge and skills.

Reporting Methods, Interpretations, and Audiences

All NAEP reporting is at the group level. NAEP reports are generally in the form of scale score distributions, with item mapping used to help describe the content meaning of various scale score points. NAEP also reports the percents of examinees at various NAEP achievement levels. Sample items are released. NAEP score reports are accompanied by general descriptions of achievement levels; such

interpretations are bolstered by the use of a large number of exercises on any one assessment.

The main basis for VNT score reporting will be to report the achievement level of an examinee's score (Below Basic, Basic, Proficient, Advanced). It is possible that probabilities that an examinee is at each of these levels could be used, or that scale scores might also be used as a means for showing how close an individual is to the next level. In any case, the representation of student performance in terms of the achievement level descriptions will be central. Thus, the key question that will affect the quality of any linkage is whether, and to what extent, descriptions based on a single form of the VNT will be congruent with the more general NAEP achievement levels descriptions.

Both NAEP and VNT reporting will likely be of interest to policy makers, educational leaders, politicians, and the American public, generally. As one or the other assessment evolves to be associated with greater stakes for students or schools, the importance of credible reporting of results can not be understated. Primary requirements of any reporting for either assessment are that reports be clear, concise, informative, technically accurate, and presented in a manner which guards, to the greatest extent possible, against likely or anticipatable misinterpretations--goals which can often conflict in their operationalization.

Available Psychometric Procedures

Currently available psychometric methods limit the options for linking NAEP and the VNT. Limitations also accrue due to the psychometric procedures already in place for producing scores on NAEP. For example, NAEP uses five correlated psychometric dimensions to represent achievement in Mathematics. A three-parameter logistic model is used to model the dichotomous items and the generalized partial credit model is used to model the constructed-response items. NAEP uses conditioning variables to improve the estimation of group-level achievement. Five plausible values sampled from the posterior distribution for each individual are intended to represent the distribution of ability for an examinee with a particular set of item responses and conditioning variables. The posterior ability distributions are collapsed to provide

percentages in each achievement level category. An item mapping procedure, done through the IRT item parameter estimates, is used to provide statements that exemplify what examinees can do who score at various scale score points.

On the other hand, it is likely that a single psychometric dimension will be used for modeling performance on each test of the VNT, at least in part because a single form of the VNT is not likely to comprise a sufficient number of items to separately model the five content dimensions. The use of a single dimension in the VNT and multiple dimensions in NAEP might lead to different psychometric dimensions being assessed. Content differences between NAEP and the VNT could also influence the psychometric dimensions that are measured. Individual level scores will be produced on the VNT; some procedure will be needed to assign examinees to achievement levels and, possibly, to indicate the probability associated with the classification of examinees with a particular estimated ability at each achievement level. It is very likely that conditioning variables will not be used in estimating VNT scores.

One approach for providing group level distributions on the VNT could involve aggregating the VNT scores over individuals. This procedure would lead to sample distributions of ability estimates as estimators of population distributions of true abilities. Because of estimation error, distributions of estimates have systematically larger variance than their population equivalents. It is anticipated that some consumers of VNT scores (e.g., policy makers, educational administrators) would be interested in the tails of the population distributions, particularly in how large the proportions of students in the Advanced or Below Basic categories are. Estimates of these proportions, based on distributions of ability estimates, would be inflated. As an alternative, population proportions could be estimated using the plausible values methodology used in NAEP--that is, as aggregates of ability estimates sampled from the posterior distributions of the students given their response vectors and scores on the NAEP conditioning variables. Although such estimates would not suffer from the bias that affects the estimator described previously, the use of a plausible values methodology would create differences in percentages of students reported to be in

various categories, depending on whether the aggregation is done on the basis of obtained VNT scores or calculated plausible values. These differences would likely create more difficulties in terms of the political milieu and performance interpretability (at all levels, local to national) than would be solved by the use of the plausible values approach. Further, this approach assumes the availability of plausible values obtained via the use of conditioning variables, although the practical problems involved in collecting the same information for the VNT as is collected for NAEP have not been explored.

Changes in the NAEP and the VNT over Time

The content of NAEP changes as conceptions of the knowledge, skills and abilities students should possess changes in the various content areas tested, and as knowledge about teaching and learning evolve. Nevertheless, compared to many other large-scale, state-level testing programs, NAEP is relatively stable. The VNT, however, is an unknown. If the evolution of the VNT mirrors that of many new testing programs, then some adjustment of content and statistical specifications will be necessary in the beginning of the program and, because of the political climate surrounding the VNT, continuing changes will be likely. In fact, it is not only the content and statistical specifications that will likely change, but all of the factors identified in Table 1 and others issues as yet unforeseen that will affect the linkage of NAEP and the VNT. For example, the composition of the student population taking the VNT likely will change, the stakes associated with the test will change, and other aspects of the administration context are likely to change. In sum, there is a high likelihood that many aspects of NAEP and the VNT will diverge over time. If this happens, then any initial linkage--even one obtained under conditions that are initially favorable--will quickly become inaccurate and inappropriate.

Another Perspective on Linking Considerations

Educational policies are often crafted in response to real or perceived crises. Unfortunately, well-

supported justifications for policy creations or educational innovations do not always accompany proposals for reforms, nor does evidence demonstrating the likelihood of success of the proposals, nor is evidence regarding the effectiveness of the proposals disseminated after the intervention has run its course (Cizek & Ramaswamy, 1999). Too typically, educational innovations or policy initiatives are introduced and implemented, obviating much impetus for studying antecedent reforms. As a result, the relationships between policy making, allocation of resources, educational innovations, and effects on key outcomes such as student achievement or educator practices remain only dimly understood.

Without question, boosting the educational achievement of all pupils is currently the focal issue in policy debates and political rhetoric regardless of which level of the American system of political subdivision one examines. All those concerned about the American educational system are struggling with how best to allocate educational resources, and the range of options receiving serious consideration, pilot testing, or full-scale implementation has, perhaps, never been broader. One recourse is remarkably common, however: large-scale, high-stakes tests are increasingly called upon to provide accurate information for informing public policy discussions or driving educational improvement initiatives. NAEP and the VNT, respectively, are perfect examples.

For an initiative to have maximum effect on improving educational achievement, however, at least two conditions are necessary: 1) the innovation must be viewed as credible, important, and desirable by the broadest possible constituency; and 2) it should have a relatively long half-life. The following sections expand upon these two ideas.

Broad Appeal

One aspect of the first condition, applied to the VNT and NAEP, means that the process of designing and developing the assessments should be based on a broad consensus of what constitutes essential knowledge and skills, and that the scoring, interpretation, and use of assessment results is

accomplished in an open, comprehensible, and fair manner. Without question, NAEP has historically attempted to withstand political and other impulses that it be molded to accomplish tasks it was not designed to, or to serve masters it was not intended to serve. The fact that NAEP has been so successful in this area is a testament to Ralph Tyler and others who foresaw the pressures that would likely affect a national report card and suggested designs that would mitigate those influences.

It remains an open question, however, whether the VNT can honor that tradition. For one thing, it is apparent that the seminal influences bringing life to the VNT are distinguishable from those that prompted the birth of NAEP. And, the extent to which NAEP has been drawn into what are--dare I say it?--ideological disputes, bodes ill with respect to the future viability and credibility of the NAEP. Let us now add to this context a new assessment--the VNT--and link it to NAEP. It seems self-evident that any embroglios which are almost certain to swirl around the development, interpretations, and uses of the higher-stakes VNT will drag NAEP into the ensuing morass. Ultimately, it is probably not too pessimistic to suggest that any damage to the credibility of the VNT will require damage control campaigns on the part of NAEP--campaigns that will be even more difficult to wage the tighter the assessments are aligned and linked.

A second aspect of breadth of appeal is the extent to which NAEP, and any derivative such as the VNT, is actually recognized and accepted by the American public. I like NAEP. Al Beaton, Governors Engler and Roemer, and Gene Johnson like NAEP, too. Unfortunately, the number of people who could even tell you what the acronym stands for is embarrassingly small once you go outside the small cadre of NAEP insiders. I asked my neighbors in the ostensibly well-informed community of Chapel Hill, North Carolina about NAEP. In conversations in which we have met our new neighbors, I introduce myself as a psychometrician who works at the University of North Carolina. With the exception of neighbors who work at Duke, so far they have been OK with my vocation. However, when I look into their eyes and tell them I have an interest in the NAEP, nearly all of them seem inclined to nervously protect the back of their necks.

The situation is, regrettably, the same with many people even in the field of education. Having

regularly taught master's level courses at the University of Toledo enrolling 40 or more teachers and principals each semester, I routinely had occasion to ask these practicing educators what they knew about NAEP. It is unusual if more than two students in a class raise their hands to indicate that they have heard of NAEP. It is typical to learn, upon further questioning of the two, that neither of them really has any idea what NAEP is, what its purpose is, or can relate any characteristic information about the assessment. Which leads to (I think) an obvious question. I apologize in advance if the following question comes off as smart-alecky, but it seems to me that it should at least be debated: "What is the public benefit of reporting the score for a student on one test he or she did take (the VNT) in terms of a plausible score on another test he or she did not take, which has no stakes, instructional value, or bearing on his or her academic success, and which virtually no one understands or cares about?"

Now, let's consider people who are knowledgeable about NAEP. I have long been interested in the inappropriate test administration practices that seem to plague high-stakes state-level pupil testing programs. I recall with particular pain hearing one teacher introduce the state assessment on the day of the test. "I'm sorry about this," she said. "We're going to have to take this test that the state *makes* us give. I wish we could spend today continuing our work on ... but we *have* to do this. Just do your best." Quite a motivational speech, eh?

Now let me quote an educator named Ernie Knoblach from Terrytown, Louisiana who is more informed than most teachers regarding NAEP, and who was not at all embarrassed to share his insights on the National Assessment with the readers of *Education Week*. In the January 12, 2000 issue, Mr. Knoblach wrote the following in a letter to the editor:

> "I read your article on the National Assessment of Educational Progress' efforts to encourage local districts to participate in this year's testing program... My experience is that the assessment is a total waste of time. We had to pull out high school students and then give the test. Many students

got up and walked out when they read it. Of the ones who stayed in the testing room, I estimate that about 90 finished within half an hour. The high school NAEP results are totally invalid and, if I were a superintendent, the test would never be given in my district."

I suppose we can take some comfort in Mr. Knoblach's use of estimation, which should please the NCTM faithful. On the other hand, given his comments, it seems to me that it is reasonable to consider whether it is a good idea to raise the profile of NAEP by linking it to the VNT. We know that many educators continue to express resistance concerning state-wide mandated pupil assessments. We also know, however, that at least educators are likely to align their instruction and efforts in order to help their students meet those external requirements. Perhaps one explanation for the success of NAEP is that it has, to some degree, been able to fly low and slow, avoiding the animus that meets mandated assessments and which find voice in comments such as those uttered by Mr. Knoblach.

But let us assume that the good ol' days of NAEP are over and that all assessments will receive even greater scrutiny and, perhaps, resistance--hardly a strong assumption. I'll give you a link. What demands to be tackled here is not a psychometric problem at all. Rather, increased attention must be paid to linking NAEP to public consciousness as a necessary indicator of American educational health, and linking NAEP to the conceptualizations that educators possess regarding obligations to their own profession and the practice of education broadly construed. I know that the National Assessment Governing Board has, in the past, commissioned work to find out how NAEP is perceived and how to broaden public understanding of the National Assessment. If NAEP is to be linked to the VNT, then such efforts aimed at broadening a consensus of understanding and support for NAEP, or any other monitoring system that is its progeny, will need to be redoubled.

It is now a recognizable phenomenon in the area of state-mandated pupil proficiency testing that score gains in the initial years of a new standard are substantially larger than those observed in succeeding years. There are many probable explanations for this phenomenon, including ceiling effects, the limits of changes in instructional alignment, initial vs. sustained motivation to meet the standard, among others. This phenomenon is examined in more detail elsewhere (see, for example, Camilli & Lugg, in press) so I will not describe it in greater length here. Let us refer to this phenomenon by the shorthand phrase from the field of economics: the law of diminishing returns.

At least one of the stated motivations for even contemplating the creation of a VNT is that it would stimulate increases in student achievement. Notably, this is different from the motivation for having NAEP, which was designed and is understood to be primarily a low-stakes monitoring system. NAEP may have endured precisely for this reason. Results on any VNTs, on the other hand, would be scrutinized for regular score gains; as the stakes increased, instructional alignment would increase, and so on. Initially at least, it is likely that large score gains would be observed, large percentages of students would move between (most of) the achievement levels, and dissatisfaction with performance would not set in until, say, the fourth or fifth testing cycle. Then what? A typical state-level response has been to "raise the bar" at this juncture. The issue then, obviously, becomes whether to: a) raise standards on a VNT (in which case there would be a concomitant degradation in the linkage between VNT and NAEP); or b) allow the VNT to languish on the plateau of "flat" student performance (in which case the investment in VNT development appears to be a great price for minimal benefit). Neither of these alternatives seems especially pleasing.

Summary

There are substantial differences between NAEP and the VNT that present serious challenges to linking the VNT to NAEP. In discussing these sorts of differences when linking tests, the *Uncommon Measures* report (National Research Council, 1999), stated that "when tests differ on any of these factors,

some limited interpretations of the linked results may be defensible while others would not" (p. 5).

The technical hurdles are serious enough. The weight of policy considerations and uncertainty about how a VNT will affect NAEP are also worth contemplating and at least speculating, as in the preceding sections, about what the likely public effects of such a linkage might be, and even whether the resources which might be devoted to the problem might be better allocated elsewhere. This point has been made previously. A policy paper produced by the Educational Testing Service Network observed that "We already have a voluntary national test; it is called the National Assessment of Educational Progress" (ETS, 1999, p. 1). In the same report, proceeding with development of the VNT was questioned: "The debate concerns the issue of doing it at all, and what benefit to American education will result" (p. 1). It is unclear to me that these issues have been addressed adequately, if at all; it occurs to me that fretting over appropriate methods for linking these assessments is analogous to straining the psychometric gnat. Please pass the insect repellant.

References

Camilli, G. A., & Lugg, C. (in press). Standard setting and accountability: Historical perspectives on the future. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Erlbaum.

Cizek, G. J., Kenney, P. A., Kolen, M. J., Peters, C. W., & van der Linden, W. J. (1999). *An investigation of the feasibility of linking scores on the proposed Voluntary National Tests and the National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board.

Cizek, G. J., & Ramaswamy, V. (1999). American educational policy: Constructing crises and crafting solutions. In G. J. Cizek (Ed.), *Handbook of educational policy* (pp. 498-519). San Diego, CA: Academic.

Educational Testing Service (ETS). (1999). Too much testing of the wrong kind; too little of the right kind in K-12 education. Report of the ETS Policy Information Center, Educational Testing Service, Princeton, NJ. [available on-line at http://www.ets.org/research/pic/testing/tmt3.html]

Holland, P. W., & Rubin, D. B. (1982). *Test equating*. New York: Academic.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Policy Information Center, Educational Testing Service. [ERIC Document Reproduction Service No. ED 353 302]

National Research Council. (1999). *Uncommon measures: Equivalency and linkage of educational tests*. Washington, DC: National Academy Press.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn, (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York: American Council on Education.

Waltman, K. K. (1997). Using performance standards to link statewide achievement results to NAEP. *Journal of Educational Measurement, 34*, 101-121.

Williams, V. S. L., Rosa, K. R., McLeod, L. D., Thissen, D., & Sanford, E. E. (1998). Projecting to the NAEP scale. Results from the North Carolina end-of-grade testing program. *Journal of Educational Measurement, 35,* 277-296.

Wise, L. L., Noeth, R. J., & Koenig, J. A. (1999). *Evaluation of the Voluntary National Tests, year 2: Final report.* Washington, DC: National Academy Press.

Table 1

Some Factors Affecting Linkages of NAEP and VNT

a) constructs measured

b) purposes of the assessments

c) administration conditions

d) stakes, visibility, motivation

e) content coverage/equivalence

f) use of achievement levels

g) reporting methods, interpretations, audiences

h) available psychometric procedures

i) changes in [a - h] over time

---

j) broad appeal

k) half life